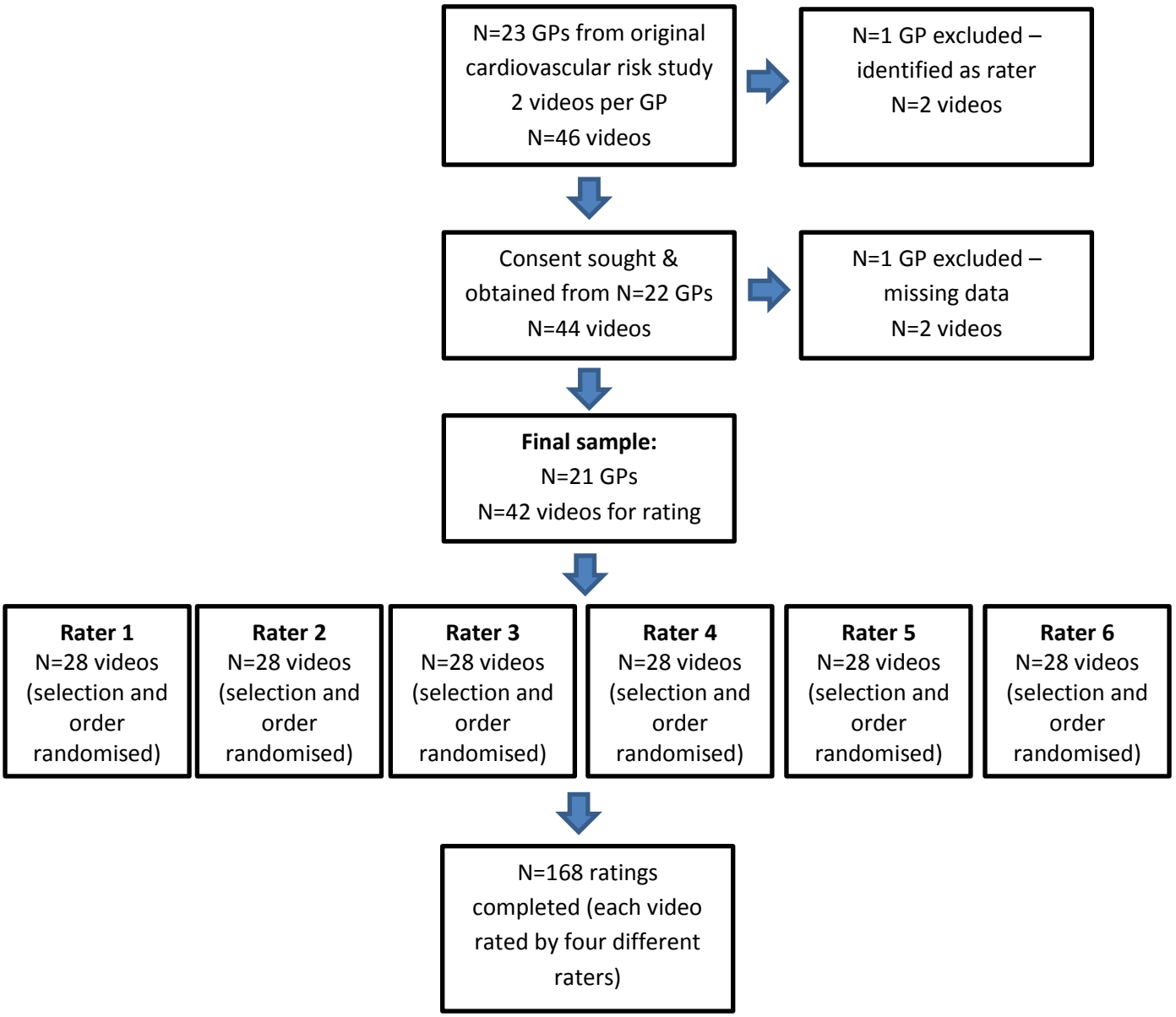
Appendix

Appendix Figure 1. Recruitment process



Randomisation of videos

In order to ensure each evaluator rated the same number of videos and that maximum overlap between evaluator was achieved we did not use simple random allocation. Instead the following procedure was used which made use of all 15 unique combinations of four raters from the total pool of six.

1. Create a dataset containing three copies of the 15 unique combinations of four raters from the total pool of six (45 entries in total). At this time raters are virtual and do not correspond to real people.
2. Delete three entries to result in the 42 combinations required. The entries deleted were chosen such that the number of videos assigned to each evaluator was reduced from 30 to 28.
3. Randomly assign a video to each entry so that now each virtual evaluator has 28 videos assigned to them for rating.
4. Within each virtual evaluator randomise the order of videos for rating.
5. Randomly assign the virtual evaluator to a real evaluator.

Distribution of videos

Each set of 28 videos were copied to an encrypted memory stick for each GP evaluator to collect. GP raters completed their name and details, and the date when they collected their data packs. They were also asked to sign a confidentiality agreement. GP raters were instructed to watch and rate the videos in numerical order. Each GP evaluator marked twenty-eight videos, and each video was marked separately by four different GP raters. All marking sheets were provided by the research team and were pre-labelled with the corresponding video identification number.

GP raters were also asked to provide written confirmation that no copies of the videos have been made during the process of data collection. All memory sticks were returned to the research team. Payment was made on completion of the work.

Estimating the reliability of the GCRS

The three level linear regression model used estimates three sources of variance;

1. between doctors (σ_d^2)
2. within doctors and between consultations (σ_c^2)
3. within consultations and between raters (σ_e^2).

Our estimate of the reliability that would be achieved for assessing single consultations (R_c) with different numbers of raters (n) is given by

$$R_c = \frac{\sigma_d^2 + \sigma_c^2}{\sigma_d^2 + \sigma_c^2 + \frac{\sigma_e^2}{n}} \qquad 1$$

Similarly our estimate of the reliability for assessing doctors performance with different numbers of raters (n) and consultations (m) per doctor is given by

$$R_d = \frac{\sigma_d^2}{\sigma_d^2 + \frac{\sigma_c^2}{m} + \frac{\sigma_e^2}{n}} \qquad 2$$

Transformation and rescaling

Because the Bland-Altman type plot revealed a clear trend of increasing variance with mean score for the untransformed data we have applied a transformation. As stated in the text a transformation based on the logit function was reasonably good at stabilising the variance. The transformation includes a rescaling such that the score is now between 0 and 10 and is given by:

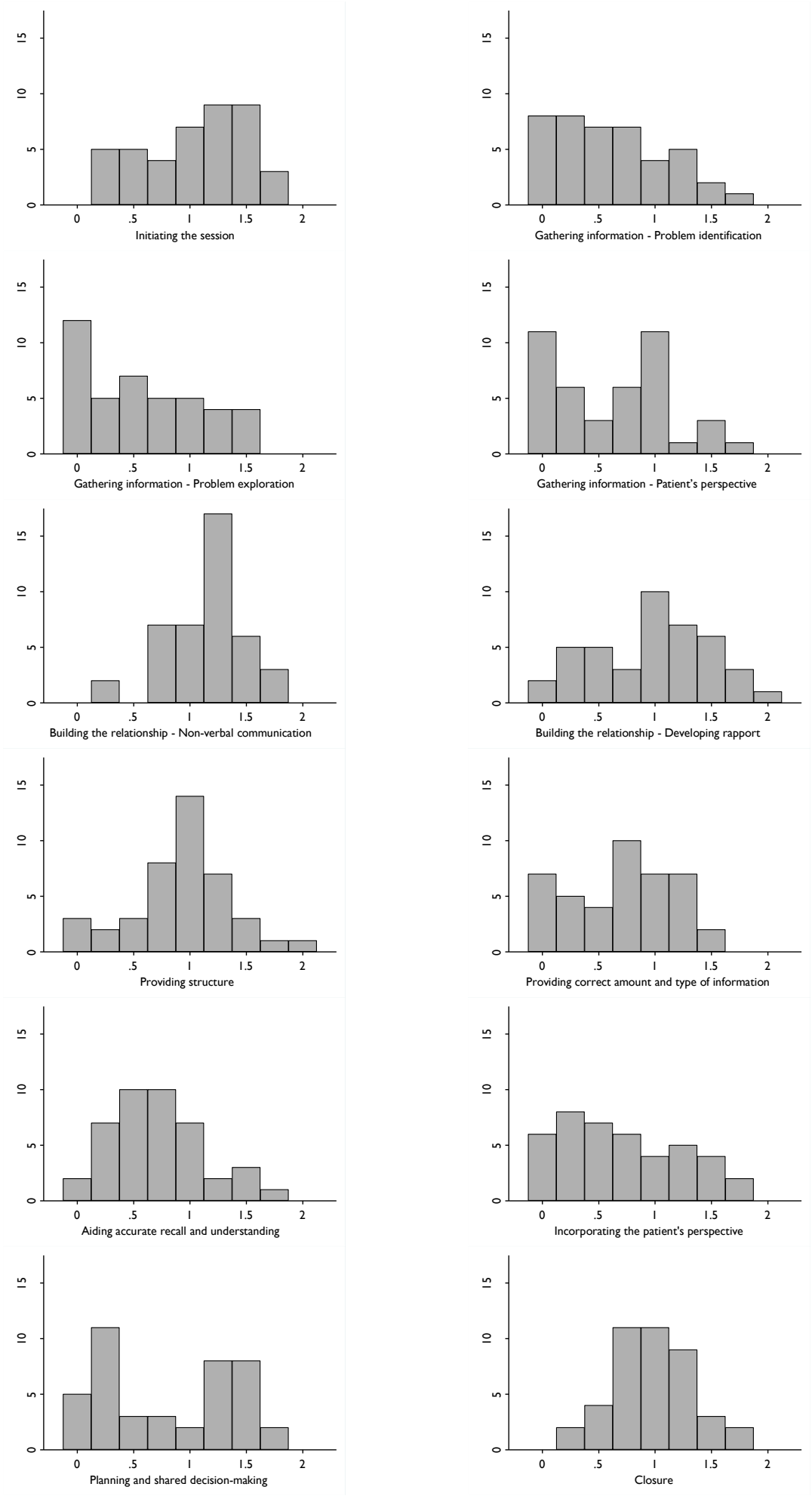
$$ts = \frac{\text{logit}(S/24) - \text{logit}(0.5/24)}{2\text{logit}(23.5/24)} \text{ if } 0 < s < 24$$
$$ts = 0 \text{ if } s = 0$$
$$ts = 10 \text{ if } s = 24 \qquad (3)$$

Alternatively the following lookup table can be used.

Appendix Table 1. Lookup table for converting an untransformed score (out of 24) to the transformed score between 0 and 10.

Original (Untransformed) Score	Transformed Score	Original (Untransformed) Score	Transformed Score
0	0.00	13	5.22
1	0.93	14	5.44
2	1.89	15	5.66
3	2.47	16	5.90
4	2.91	17	6.15
5	3.27	18	6.43
6	3.57	19	6.73
7	3.85	20	7.09
8	4.10	21	7.53
9	4.34	22	8.11
10	4.56	23	9.07
11	4.78	24	10.00
12	5.00		

Appendix Figure 2 - Histograms showing the distribution of mean consultation scores for each domain of the GCRS (possible values 0 to 2)



Appendix Table 2 – Crude reliability of each domain of the GCRS when used for evaluating consultations using different numbers of raters.

Number of raters	Initiating the session	Gathering information			Building the relationship		Providing structure	Providing correct amount and type of information	Aiding accurate recall and understanding	Incorporating the patient's perspective	Planning and shared decision-making	Closure
		Problem identification	Problem exploration	Patient's perspective	Non-verbal communication	Developing rapport						
1	0.30 (0.15, 0.36)	0.45 (0.30, 0.52)	0.41 (0.24, 0.49)	0.35 (0.21, 0.42)	0.17 (0.02, 0.21)	0.41 (0.23, 0.51)	0.33 (0.10, 0.45)	0.28 (0.10, 0.36)	0.27 (0.09, 0.36)	0.41 (0.19, 0.48)	0.46 (0.32, 0.51)	0.14 (0.05, 0.18)
2	0.46 (0.26, 0.53)	0.62 (0.46, 0.69)	0.58 (0.39, 0.66)	0.52 (0.34, 0.59)	0.28 (0.04, 0.35)	0.58 (0.38, 0.67)	0.49 (0.18, 0.62)	0.43 (0.18, 0.53)	0.42 (0.16, 0.53)	0.58 (0.32, 0.65)	0.63 (0.48, 0.68)	0.25 (0.10, 0.31)
3	0.56 (0.35, 0.63)	0.71 (0.56, 0.77)	0.68 (0.49, 0.75)	0.62 (0.44, 0.69)	0.37 (0.05, 0.45)	0.68 (0.48, 0.75)	0.59 (0.25, 0.71)	0.53 (0.25, 0.63)	0.52 (0.22, 0.63)	0.67 (0.42, 0.73)	0.72 (0.58, 0.76)	0.33 (0.14, 0.40)
4	0.63 (0.41, 0.69)	0.76 (0.63, 0.81)	0.74 (0.56, 0.80)	0.68 (0.51, 0.75)	0.44 (0.07, 0.52)	0.74 (0.55, 0.80)	0.66 (0.31, 0.77)	0.60 (0.31, 0.69)	0.59 (0.28, 0.69)	0.73 (0.49, 0.79)	0.77 (0.65, 0.81)	0.40 (0.18, 0.47)
5	0.68 (0.47, 0.74)	0.80 (0.68, 0.85)	0.78 (0.62, 0.83)	0.73 (0.57, 0.79)	0.50 (0.09, 0.58)	0.78 (0.60, 0.84)	0.71 (0.36, 0.81)	0.66 (0.36, 0.74)	0.64 (0.32, 0.74)	0.77 (0.55, 0.82)	0.81 (0.70, 0.84)	0.45 (0.22, 0.53)
6	0.72 (0.51, 0.77)	0.83 (0.72, 0.87)	0.81 (0.66, 0.85)	0.76 (0.61, 0.81)	0.54 (0.10, 0.62)	0.81 (0.64, 0.86)	0.75 (0.40, 0.83)	0.70 (0.40, 0.77)	0.68 (0.36, 0.77)	0.80 (0.59, 0.85)	0.83 (0.73, 0.86)	0.50 (0.25, 0.57)
7	0.75 (0.55, 0.80)	0.85 (0.75, 0.88)	0.83 (0.69, 0.87)	0.79 (0.65, 0.84)	0.58 (0.12, 0.66)	0.83 (0.68, 0.88)	0.77 (0.44, 0.85)	0.73 (0.44, 0.80)	0.72 (0.40, 0.80)	0.83 (0.63, 0.87)	0.85 (0.76, 0.88)	0.54 (0.28, 0.61)
8	0.77 (0.59, 0.82)	0.87 (0.77, 0.90)	0.85 (0.72, 0.89)	0.81 (0.68, 0.85)	0.61 (0.13, 0.69)	0.85 (0.71, 0.89)	0.80 (0.47, 0.87)	0.75 (0.47, 0.82)	0.74 (0.43, 0.82)	0.85 (0.66, 0.88)	0.87 (0.79, 0.89)	0.57 (0.31, 0.64)
9	0.79 (0.61, 0.83)	0.88 (0.79, 0.91)	0.86 (0.74, 0.90)	0.83 (0.70, 0.87)	0.64 (0.15, 0.71)	0.86 (0.73, 0.90)	0.81 (0.50, 0.88)	0.77 (0.50, 0.83)	0.77 (0.46, 0.84)	0.86 (0.68, 0.89)	0.88 (0.81, 0.90)	0.60 (0.34, 0.67)
10	0.81 (0.64, 0.85)	0.89 (0.81, 0.92)	0.88 (0.76, 0.91)	0.84 (0.72, 0.88)	0.67 (0.16, 0.73)	0.87 (0.75, 0.91)	0.83 (0.53, 0.89)	0.79 (0.53, 0.85)	0.78 (0.49, 0.85)	0.87 (0.71, 0.90)	0.89 (0.82, 0.91)	0.62 (0.36, 0.69)